Generalizability assessment of head and neck cancer NTCP models based on the TRIPOD criteria

Sharabiani, Marjan ; Clementel, Enrico ; Andratschke, Nicolaus ; Hurkmans, Coen

Abstract: This review aimed to provide an overview of the level of maturity of normal tissue complication probability (NTCP) models for head and neck cancer (HNC) patients. A systematic literature review was performed to retrieve NTCP models for HNC toxicities. Patient population characteristics, NTCP model and the predictors, treatment technique and endpoint definition were extracted per article. Models were then scored based on the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) consensus guidelines to evaluate their generalizability. 335 articles on photon and proton therapy of HNC were identified and 52 relevant articles were further analyzed. Eighteen articles on xerostomia and sticky saliva (TRIPOD types 1a-2b: 15; TRIPOD type 3: 1; TRIPOD types 4a: 1 4b:1), thirteen articles on dysphagia and tube feeding dependence (TRIPOD types 1a-2b: 7; TRIPOD type 3: 2; TRIPOD types 4a:2 4b:2), five articles on oral mucositis (TRIPOD types 1a-2b: 4; TRIPOD type 4b: 1), seven articles on hypothyroidism (TRIPOD types 1a-2b: 4; TRIPOD type 3: 1; TRIPOD types 4a: 1 4b:1), four articles on hearing loss and tinnitus (TRIPOD type 1a: 4) and ten articles on esophagitis (TRIPOD types 1a-2b: 9; TRIPOD type 4a: 1) were included. External validation studies of HNC NTCP models are scarce. Moreover, the majority of them were validating a model developed by the same researchers. Only 2 independent external validation studies were found. There is a strong need to publish external validation studies to get more mature NTCP models applicable in clinical practice.

DOI: https://doi.org/10.1016/j.radonc.2020.02.013

Radiotherapy and Oncology 146 (2020) 143–150

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Systematic Review

Generalizability assessment of head and neck cancer NTCP models based on the TRIPOD criteria

Marjan Sharabiani [a,*], Enrico Clementel [a], Nicolaus Andratschke [b,d], Coen Hurkmans [c,d]

[a] European Organisation for Research and Treatment of Cancer (EORTC) Headquarters, Brussels, Belgium; [b] Department of Radiation Oncology, University Hospital Zürich, University of Zurich, Zürich, Switzerland; [c] Department of Radiation Oncology, Catharina Hospital, Eindhoven, The Netherlands; and [d] EORTC Radiation Oncology Group, Brussels, Belgium

ARTICLE INFO

Article history:
Received 30 October 2019
Received in revised form 6 February 2020
Accepted 17 February 2020
Available online 7 March 2020

Keywords:
NTCP models
Head and neck cancer
TRIPOD type

ABSTRACT

This review aimed to provide an overview of the level of maturity of normal tissue complication probability (NTCP) models for head and neck cancer (HNC) patients. A systematic literature review was performed to retrieve NTCP models for HNC toxicities. Patient population characteristics, NTCP model and the predictors, treatment technique and endpoint definition were extracted per article. Models were then scored based on the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) consensus guidelines to evaluate their generalizability. 335 articles on photon and proton therapy of HNC were identified and 52 relevant articles were further analyzed. Eighteen articles on xerostomia and sticky saliva (TRIPOD types 1a-2b: 15; TRIPOD type 3: 1; TRIPOD types 4a: 1 & 4b:1), thirteen articles on dysphagia and tube feeding dependence (TRIPOD types 1a-2b: 7; TRIPOD type 3: 2; TRIPOD types 4a:2 & 4b:2), five articles on oral mucositis (TRIPOD types 1a-2b: 4; TRIPOD type 4b: 1), seven articles on hypothyroidism (TRIPOD types 1a-2b: 4; TRIPOD type 3: 1; TRIPOD types 4a: 1 & 4b:1), four articles on hearing loss and tinnitus (TRIPOD type 1a: 4) and ten articles on esophagitis (TRIPOD types 1a-2b: 9; TRIPOD type 4a: 1) were included. External validation studies of HNC NTCP models are scarce. Moreover, the majority of them were validating a model developed by the same researchers. Only 2 independent external validation studies were found. There is a strong need to publish external validation studies to get more mature NTCP models applicable in clinical practice.

Treatment planning of HNC patients often represents a challenge. The complexity of HNC treatment planning is due to the close proximity of organs at risk (OARs) to the planning target volume (PTV). Multiple guidelines are available in the literature on delineation of target volumes and OARs, including dose objectives and constraints [1–3]. If simultaneous sparing of OAR and full PTV coverage is not possible, physicists and physicians adopt strategies to either spare healthy tissue at the cost of non-uniform irradiation of the PTV or irradiate above accepted dose constraints to maintain target coverage, thereby accepting higher risk of complications.

Several NTCP models have become available in the literature in the past years for a number of different organs and endpoints. Potentially, NTCP models can help quantify individual risks to develop specific toxicities and help physicians and physicists make an informed decision on a personalized treatment strategy for the patient. Furthermore, NTCP models could be used as quality indicators, in the frame of semi-automated quality assurance (QA) of treatment plans.

NTCP models are often developed and internally validated either by cross-validation, bootstrapping or by randomly splitting original dataset in test and validation sets [4]. However, validation in an independent dataset, i.e. evaluation of a model's performance beyond its training set, is a crucial step before clinical implementation of a prediction model in the clinic. Prediction model performance measures were traditionally defined as the overall model performance, discriminative ability and calibration [5]. Overall model performance is normally indicated by the Brier score, while discriminative ability is specified by the concordance (c) statistics or the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, and calibration is explained by the goodness-of-fit statistics [5]. The Brier score represents the mean squared error between the actual outcomes, Y, and predictions, p. The Brier score ranges from 0 for a perfect model to 0.25 for a non-informative model, assuming that the incidence of the outcome is 50%. The lower the outcome incidence, the lower the maximum threshold for a non-informative model [5].

The TRIPOD statement [6] is an annotated checklist of items created to enhance the quality and transparency of reporting prediction models' development, validation and performance. Four items in the TRIPOD statement in particular address the validation

* Corresponding author at: EORTC Headquarters, Av. E. Mounier 83/11, 1200 Brussels, Belgium.
E-mail address: marjan.sharabiani@eortc.org (M. Sharabiani).

https://doi.org/10.1016/j.radonc.2020.02.013
0167-8140/

of models; depending on certain factors a model can be classified in 4 main types and 3 subtypes. Although not stated in the original TRIPOD statement, in this article, we split TRIPOD Type 4 studies into 2 subtypes based on the level of independence from the original investigators who developed the model.

- Type 1a defines a model where the predictive performance is directly evaluated using exactly the same dataset used in model development.
- Type 1b defines a model where performance is evaluated on a dataset obtained using resampling techniques on the development dataset (internal validation).
- Type 2a defines a model where a dataset is randomly split into two groups, one used to develop a model and the other to evaluate its predictive performance.
- Type 2b applies a more robust technique by non-random splitting of data (by location, time etc.).
- Type 3 defines a model developed and evaluated on separate datasets by the model developers, for example using data from another institution.
- Type 4a defines a published model which is externally validated by the same investigators who developed the model.
- Type 4b defines a published model which is externally validated by independent investigators.

Brodin et al. [7] performed a comprehensive review of NTCP models in HNC focusing on the validity of QUANTEC dose constraints and variation of statistical methodology and reporting of various NTCP models. In contrast, our review was carried out to give an overview of the currently published NTCP models and their respective level of external validation as defined by the TRIPOD statement [6]. The aim is to provide a perspective of the most generalized, robust and transportable models currently available and their potential application both in the clinic and in clinical trials.

Materials and methods

A PubMed search was carried out in March 2019 and updated through August 2019. Keywords used were combinations of "NTCP" OR "normal tissue complication probability" OR "dose–response" AND "radiotherapy" OR "radiation" OR "radiation induced" OR "complication" for each of the following relevant head and neck radiation-induced toxicities: "xerostomia" OR "dry mouth"; "sticky saliva"; "dysphagia" OR "swallowing dysfunction"; "esophagitis"; "oral mucositis"; "hypothyroidism"; "hearing loss" and "tinnitus". The initial search resulted in 213 articles. No limit to the earliest publication date was applied. The search was extended by references cited in the retrieved articles.

The literature search was extended for NTCP models either developed or validated for proton beam therapy of HNC, using combinations of "NTCP" OR "Normal Tissue Complication Probability" AND "protons". Overall 122 references were retrieved and analyzed.

Articles in which a quantitative dose–response model was developed and/or validated were finally selected for further analysis. Applying our selection criteria, 52 full length articles were selected from the initial search results and analyzed. The vast majority (51 articles) described development and/or validation of models for patients treated with photons. One article reported on external validation of NTCP models developed for photon therapy on a patient cohort treated with proton beam therapy [8].

After full text review, data extraction was undertaken. Items for extraction included: manuscript identifiers (author, title), patient population characteristics, type of the NTCP model and predictive variables constituting the model (dosimetric and clinical), radiotherapy (RT) delivery technique and fractionation schedule, chemotherapy agent/schedule (if available), endpoint(s) definition as well as the time point for toxicity measurement and toxicity scoring system and finally model's performance measures included in the paper. The model described in each article was then scored based on the TRIPOD statement on model development and validation.

Results

Out of the initial pool of 52 articles, 80% reported exclusively on model development, i.e. TRIPOD types 1a, 1b, 2a or 2b; 7% reported on development and validation, i.e. TRIPOD type 3 and 13% reported on validation of a published model, i.e. TRIPOD types 4a or 4b. According to our analysis, Blanchard's article [8] on external validation of five different toxicities (oral mucositis, dry mouth, dysphagia, tube feeding dependence and hypothyroidism) from various source articles as well as external validation of physician-rated dysphagia by Hansen et al. [9] using patients in the randomized controlled DAHANCA19 trial as the validation cohort, were the only independent external validation articles (Type 4b). All other type 4 studies were validated on separate datasets by the same authors who developed the model (Type 4a).

Xerostomia and sticky saliva

Xerostomia and sticky saliva are the most common complications after RT of HNC. This is reflected in the high number of results in our search, with a total number of 18 papers referring to 16 dedicated models. Ten articles referred to Lyman-Kutcher-Burman (LKB) type models, while the remaining were multivariate logistic regression NTCP models. Parotid gland mean dose and the mean dose to submandibular and sublingual glands were suggested as the most common predictive factors for xerostomia and sticky saliva incidence respectively.

A summary of the salient characteristics of each model in each paper is reported in Table 1. Comprehensive information is available in Supplementary Table S1.

Of 18 studies, 14 models fall in to the TRIPOD type 1a or 1b, from which only 2 articles reported on overall model performance either by means of R^2 alone [21] or R^2 in combination with the scaled Brier score [24]. Eight of these articles reported discriminative performance by means of AUC, while calibration was only measured in one article [14] using calibration slope. Five of these articles did not report any performance measures.

Only 1 model [19] used non-random splitting of data to develop a model and evaluate its optimism according to two different time points, i.e. TRIPOD type 2b. Discriminative power of the model was the only reported performance measure by AUC.

One article was also dedicated to model development and validation by the same research group, but using two separate datasets, i.e. TRIPOD type 3 [16]. AUC was the single reported performance measure. This study also provided evidence that the predictive model developed for head and neck squamous cell carcinoma (HNSCC) patients did not perform well among patients with nasopharyngeal cancer (NPC) and vice versa.

External validation was only performed in 2 articles for 2 logistic regression models, the only TRIPOD type 4 analysis was found for xerostomia [13,8]. In the study by Beetz et al. [13], overall model performance, discrimination and calibration were reported by Brier score, AUC and Hosmer-Lemeshow goodness-of-fit test respectively. The paper also showed that models developed for patients treated with 3D-CRT, could not be generalized to patients treated with IMRT without external validation.

Table 1
NTCP models of xerostomia.

Endpoint	Dosimetric risk factors	Clinical risk factors	Type	Performance Measures	Reference
Salivary excretion function (scintigraphy)	Parotid Gland mean dose	None	1a	None	[10]
Grade 3 + xerostomia (scintigraphy)	Parotid gland mean dose	None	1a	AUC	[11]
Patient-rated moderate to severe xerostomia	Parotid Gland mean dose	Age, baseline xerostomia score	1b	AUC	[12]
Patient-rated moderate to severe sticky saliva	Submandibular and sublingual glands mean dose	Age, baseline sticky saliva score			
Patient-rated moderate to severe xerostomia	Parotid gland mean dose	Age, baseline xerostomia score	4a	AUC, R^2 Nagelkerke, Brier score, Hosmer-Lemeshow goodness of fit test	[13]: External validation of [12]
Patient-rated moderate to severe sticky saliva	Submandibular and sublingual glands mean dose	Age, baseline sticky saliva score			
Patient-rated moderate to severe xerostomia	Contralateral parotid gland mean dose	Baseline xerostomia score	1b	AUC, discrimination slope, calibration slope	[14]
Patient-rated moderate to severe sticky saliva	Contralateral submandibular, sublingual and soft palate glands mean dose	None			
Patient-rated moderate to severe sticky saliva	Contralateral submandibular, sublingual and soft palate glands mean dose	None	4b	AUC, Hosmer-Lemeshow goodness of fit test	[8]: External validation of [14]
Patient-rated moderate to severe xerostomia (3 months' time point)	Ipsi/contralateral parotid glands mean dose	Age	1b	AUC	[15]
Patient-rated moderate to severe xerostomia (12 months' time point)	Contralateral and ipsilateral parotid glands mean dose	Smoking, education, T-stage			
Patient-rated moderate to severe xerostomia (3 months' time point & HNSCC)	Ipsi/contralateral parotid glands mean dose	Age	3	AUC	[16]
Patient-rated moderate to severe xerostomia (3 months' time point & NPC)	Contralateral parotid gland mean dose	Financial status, Age			
Patient-rated moderate to severe xerostomia (12 months' time point & HNSCC)	Ipsi/contralateral parotid glands mean dose	T-stage			
Patient-rated moderate to severe xerostomia (12 months' time point & NPC)	Ipsi/contralateral parotid glands mean dose	Education			
Grade 4 xerostomia (salivary flow)	Parotid Gland mean dose	None	1a	None	[17]
Xerostomia (25% salivary flow reduction)	Parotid Gland mean dose	None	1b	AUC	[18]
Grade 1 + and grade 2 + xerostomia	Parotid Glands mean dose	None	2b	AUC	[19]
Grade 3 xerostomia	Parotid Gland mean dose	None	1a	AUC	[20]
Acute (CTCAE v3.0) and chronic (RTOG/EORTC and LENTSOMA) toxicity	Parotid Gland mean dose	None	1a	R^2	[21]
Stimulated parotid flow	Parotid Gland mean dose	None	1a	None	[22]
Stimulated parotid flow	Parotid Gland mean dose	None	1a	None	[23]
Patient-rated moderate to severe xerostomia	Ipsi/contralateral submandibular parotid and contralateral submandibular gland, oral cavity mean dose	Baseline xerostomia, Age, T-stage	1a	Scaled Brier score, R^2 Nagelkerke, AUC	[24]
Patient-rated xerostomia	Combined contralateral (parotid and submandibular) glands mean dose	None	1a	AUC	[25]
Stimulated parotid flow ratio <25% of the pretreatment flow rate (grade 4 xerostomia based on RTOG/EORTC)	Parotid Gland mean dose	None	1a	None	[26]

Finally, Blanchard et al. [8] used AUC to report discrimination and Hosmer-Lemeshow goodness-of-fit test to report calibration.

In summary, 15 of the aforementioned articles (10 LKB and 5 logistic regression models) have merely developed a predictive model (TRIPOD types 1a, 1b & 2b), while only one article considered model development in combination with validation (TRIPOD type 3). Two articles on xerostomia were scored as TRIPOD types 4. However, only in one case [8] the validation could be scored as type 4b. This paper, however, evaluated the performance of a model developed for photon therapy on a dataset of patients treated with protons.

Dysphagia/swallowing dysfunction and tube feeding dependence

We observed a total of 13 articles (9 distinct models) on dysphagia/tube feeding dependence, summarized in Table 2. Nine out of 13 articles used multivariate logistic regression model, indicating that physician and patient-rated swallowing dysfunction in HNC patients could not be described by mere dose distribution parameters. It is apparent from the data provided in Table 2 that there is a large heterogeneity in dosimetric risk factors among NTCP models of dysphagia. Nonetheless, pharyngeal constrictor muscle and supraglottic larynx mean dose were commonly found to correlate with the risk of dysphagia. Detailed information on each article can be found in Supplementary Table S2.

Of 13 articles, two studies were assigned TRIPOD type 1a; of these, one used AUC as the only predictive performance measure [25], while no model performance was evaluated in the other study [32]. Four articles were assigned TRIPOD type 1b; 3 of which performed calibration and discrimination evaluation by measures of AUC, discrimination value or calibration slope while one study only reported AUC [27]. TRIPOD type 2b was assigned to 1 study [4], in which by utilizing a multivariate logistic regression model, the authors assessed model validation temporally, predicting

Table 2
NTCP models of dysphagia/swallowing dysfunction and tube feeding dependence.

Endpoint	Dosimetric risk factors	Clinical risk factors	Type	Performance measures	Reference
Grade 2–4 swallowing dysfunction	Superior pharyngeal constrictor muscle & supraglottic larynx mean dose	None	1b	AUC	[27]
Patient-rated moderate to severe swallowing complaints	Supraglottic larynx, middle and superior pharyngeal constrictor muscles mean dose, V60 of the oesophageal inlet muscle	RT technique, age, tumor site			
Grade 2–4 swallowing dysfunction	Superior pharyngeal constrictor muscle & supraglottic larynx mean dose	None	4a	Scaled Brier score, AUC, discrimination slope, Hosmer-Lemeshow test	[28]: External validation of [27]
Grade 2–4 physician-rated dysphagia	Same as above	None	4b	Brier score, AUC, Calibration plot	[9]: External validation of [27]
Grade 2–4 swallowing dysfunction	Superior pharyngeal constrictor muscle & supraglottic larynx mean dose	None	4b	AUC, Hosmer-Lemeshow test	[8]: External validation of [27]
Grade 2 + swallowing dysfunction	Pharyngeal muscles mean dose	T-stage, bilateral neck irradiation, weight loss, primary tumor site, treatment modality	2b	None	[4]
Grade 3 + and less than grade 3 dysphagia	V80-V100 of pharyngeal mucosa	None	3	AUC	[29]
Grade 3 + dysphagia	Pharyngeal mucosa mean dose	None	1a	AUC, calibration slope, discrimination value	[30]
Grade 3 + dysphagia	V45 of the cervical esophagus, Cricopharyngeal muscle mean dose	None	1b	AUC, Calibration slope, Calibration intercept, Discrimination value	[31]
Grade 2 + dysphagia	Total constrictor muscle mean dose	Disease site, Age	1a	None	[32]
Grade 2 + dysphagia	superior pharyngeal constrictors mean dose	None	1a	AUC	[25]
Tube feeding dependence	Superior/Inferior pharyngeal constrictor muscle, contralateral parotid, cricopharyngeal muscle mean dose	Advanced T-stage, moderate weight loss, severe weight loss, accelerated RT, chemoradiation, radiotherapy plus cetuximab	1b	AUC, discrimination slope, Hosmer-Lemeshow chi square test	[33]
Tube feeding dependence	Same as above	Same as above	4a	Nagelkerke's R^2, AUC, discrimination slope	[34]: External validation of [33]
Tube feeding dependence	Same as above	Same as above	4b	AUC, Hosmer Lemeshow test	[8]: External validation of [33]
Tube feeding dependence	None	T-stage, N-stage, moderate weight loss, bilateral neck irradiation, accelerated RT, Chemo-RT	3	AUC, Discrimination slope, Hosmer-Lemeshow chi square test	[35]

swallowing dysfunction at later time points. However, there was no quantitative measures on model performance.

Studies by Wopken et al. [35] and Dean et al. [29] were assigned TRIPOD type 3 in which multivariate logistic regression with penalized learning method and penalized logistic regression models were used respectively. High values of AUC in training (AUC = 0.86) as well as validation cohorts (AUC = 0.82) in model by Wopken et al. [35] were representative of model's good discriminative ability. Penalized logistic regression model in the article by Dean et al. [29] was also selected based on its best performance (assessed by AUC) compared to support vector classification and random forest classification (RFC) models.

We identified four studies with TRIPOD types 4a & 4b. AUC was reported in all studies, while overall model performance was reported in 3 articles out of 4, in one by Nagelkerke's R^2 [34], in the other by the scaled Brier score [28] and in [9] by Brier score. Calibration was reported in 3 out of 4 studies by measures of Hosmer-Lemeshow test [28,8] and visualized by calibration plot in [9].

In summary, from the 13 articles on dysphagia and tube feeding dependence, 9 articles were restricted to model development, either alone (TRIPOD type 1a, 1b or 2b) or in combination with model validation (TRIPOD type 3), while only 4 articles performed model validation (TRIPOD types 4a & 4b). The multivariate model on tube feeding dependence at 6 months post-treatment developed by Wopken et al. [33] was validated in an external cohort by the same group [34] (TRIPOD type 4a). External validation of this model was also performed by a totally independent patient cohort treated by proton beam therapy by Blanchard et al. [8]. Clinical validation of the multivariate logistic regression model performed by Christianen et al. [28] was also scored as having TRIPOD type 4a, although it was also an effort to clinically validate the model previously published by the same group. The multivariate model for physician-rated dysphagia developed by Christianen et al. [27] was independently validated by Hansen et al. [9]. The results of the closed testing procedure indicated that an intercept refitting of the original model creates a better fit for the validation cohort. To our knowledge, among the aforementioned 4 studies, the studies by Blanchard et al. [8] and Hansen et al. [9] were the only TRIPOD type 4b studies.

Oral mucositis

We found 5 articles on oral mucositis. Diverse risk factors were predicted by different studies. Penalized logistic regression and Random Forest Classification methods were used by Dean et al. [36,37] in two publications, while LKB [38] and Logit models [30] were used as well for modeling of oral mucositis. The summary of papers could be find in Table 3, and detailed information is in the Supplementary Table S3.

Table 3
NTCP models of oral mucositis.

Endpoint	Dosimetric risk factors	Clinical risk factors	Type	Performance Measures	Reference
Severe (grade 3+) and non-severe (less than grade 3) mucositis	V180 of the oral cavity/oral mucosa	Primary disease site	1b	AUC, Brier score, Calibration slope & intercept	[36]
Severe (grade 3 +) and non-severe (less than grade 3) mucositis	V180 & V220 of the oral cavity/ oral mucosa	Age	1b	AUC	[37]
Grade 3 oral mucositis	Mean dose to the oral mucosa	None	1a	R^2 value for goodness of fit	[30]
Grade 3 + acute mucositis	Same as above	Same as above	4b	R^2, AUC, Hosmer-Lemeshow test	[8]: External validation of [30]
Grade 3 + acute mucositis	Mean dose to oral/pharyngeal mucosa	None	1a	AUC	[38]

Of 5 articles, 4 models described model development (TRIPOD type 1a & 1b), while only 1 article validated a previously published model [8] (TRIPOD type 4b).

Four of the articles reported AUC discrimination measures. Only 1 study with TRIPOID type1b [36] performed all model performance measures, i.e. overall model performance was reported by Brier score, discrimination power by AUC and calibration was presented by calibration slope and intercept, while the other TRIPOD 1b study [37] only reported AUC as the discriminative power performance. In the article by Bhide et al. [30] (TRIPOD type 1a) the only performance measure reported was R^2, while in the corresponding independent external validation study by Blanchard et al. [8] on proton therapy patients reported both discrimination ability and calibration of the model by AUC and Hosmer-Lemeshow test respectively. AUC was the only reported performance measure for TRIPOD 1a study by [38].

Despite the relatively good performance of these models assessed by cross-validation techniques in the internal dataset, the predictive power of these models has been rarely validated on an external dataset to confirm the generalizability of the predictions. To the best of our knowledge, the external validation study performed by Blanchard et al. [8] was the only study which could be assigned TRIPOD type 4b.

Hearing loss and tinnitus

One article reported on the incidence of tinnitus after HNC IMRT [39] and three articles addressed hearing loss are summarized in Table 4. Detailed information is reported in Supplementary Table S4. Radiation dose to cochlea is the most reported major predictive factor of hearing loss. LKB model was used in all four articles, however in the study by Cheraghi et al. [40] logistic, relative seriality, Critical volume individual and Critical volume population models were used as well. LKB and logistic models were used in the model developed on tinnitus [39].

None of the studies found on hearing loss and tinnitus were validated internally (TRIPOD type 1a). Considering performance measures of NTCP models on hearing loss: AUC was reported in the model developed by De Marzi et al. [41] and calibration was reported for the six models developed by Cheraghi et al. [40], while no performance measures was used by Mosleh-Shirazi et al [42].

The overall performance, discrimination and calibration of the LKB and logistic NTCP models on tinnitus [39] were measured using the scaled Brier score and AUC, Hosmer-Lemeshow tests and the calibration slope respectively which showed satisfactory results with similar performance for the two models.

Hypothyroidism

We identified 7 articles (5 models) with a quantitative analysis of thyroid gland radiation-induced complications, summarized in Table 5. A detailed description can be found in Supplementary Table S5.

Multivariate logistic model was used for all articles. Mean dose to thyroid gland and thyroid volume were the most prominent factors associated with the risk of hypothyroidism.

Out of 7 studies, 4 models had TRIPOD types 1a or 1b. One study which was assigned as having TRIPOD type 1a [43], used 2-sample Kolmogorov-Smirnov goodness-of-fit test to evaluate calibration, while no other performance measures was performed for discrimination or overall performance. Discrimination was measured by AUC and discrimination slope by Boomsma et al. [44] (TRIPOD type 1b). In the other TRIPOD type 1b study by Ronjom et al. [46] no performance measure was performed.

One study developed a biochemical hypothyroidism model and validated it in a different dataset (TRIPOD type 3) [45]. Discriminative performance was the only reported performance measure by AUC in this study.

Two studies were assigned TRIPOD types 4a and 4b. One, type 4b external validation, used a model originally developed for patients treated with 3D-CRT or IMRT [44] and validated them for patients treated with intensity modulated proton beam therapy [8]. Calibration and discrimination were measured by the Hosmer-Lemeshow test and AUC respectively. The other TRIPOD 4a study

Table 4
NTCP models of hearing loss and tinnitus.

Endpoint	Dosimetric risk factors	Clinical risk factors	Type	Performance Measures	Reference
Grade 2 + tinnitus	Mean dose to cochlea	None	1a	Scaled Brier score, AUC, calibration slope	[39]
≥Grade 1–2 ear and pituitary gland late complications	Mean dose and gEUD with a values ≤1.2 to the cochlea	None	1a	AUC	[41]
Sensorineural hearing loss	Mean dose to cochlea	None	1a	2-sample Kolmogorov-Smirnov goodness of fit test	[40]
Conductive hearing loss	Mean dose to the middle ear	None	1a	None	[42]

Table 5
NTCP models of hypothyroidism.

Endpoint	Dosimetric risk factors	Clinical risk factors	Type	Performance measures	Reference
Grade 1 + hypothyroidism	Mean dose to thyroid gland	None	1a	2-sample Kolmogorov-Smirnov goodness of fit test	[43]
Subclinical or clinical hypothyroidism	Mean dose to thyroid gland	Thyroid gland volume	1b	AUC, discrimination slope	[44]
Hypothyroidism	Same as above	Same as above	4b	AUC, Hosmer-Lemeshow test	[8]: External validation of [44]
Biochemical hypothyroidism (stimulated TSH*)	V30 of thyroid gland	Thyroid gland volume, gender	3	AUC	[45]
Biochemical hypothyroidism (stimulated TSH)	Mean dose to thyroid gland	Thyroid gland volume, latency	1b	None	[46]
Biochemical hypothyroidism (stimulated TSH)	Same as above	Same as above	4a	Pearson's R^2	[47]: External validation of [46]
Biochemical hypothyroidism (stimulated TSH)	V50 Gy & the maximum dose of the pituitary gland	Gender, chemotherapy	1b	Hosmer-Lemeshow test, AUC	[48]

*Thyroid stimulating hormone.

Table 6
NTCP models of esophagitis.

Endpoint	Dosimetric risk factors	Clinical risk factors	Type	Performance Measures	Reference
Grade ≥2 and grade ≥3 (acute esophageal toxicity)	V50 of the esophagus	None	1a	None	[49]
Grade ≥2 acute esophageal toxicity	Esophagus mean dose	tumor stage, gender, concurrent chemotherapy	1b	AUC, Calibration plot, Hosmer Lemeshow test, Nagelkerke's R^2	[50]
Grade ≥2 acute esophageal toxicity	Same as above	Same as above	4a	AUC	[51]: External validation of [50]
Grade ≥2 acute esophagitis	Esophagus mean dose	Concurrent Chemotherapy	1b	AUC, Calibration plot	[52]
Grade >2 acute esophagitis	Esophagus mean dose	None	1a	None	[53]
Grade 2 >esophagitis	Esophagus mean dose	None	1a	None	[54]
Maximal acute esophageal toxicity at any time point	Esophagus mean dose, V35 of esophagus	Concurrent chemo-RT	1a	None	[55]
Grade 2 >acute esophagitis	V38 of esophagus, Esophagus mean dose	None	1a	None	[56]
Patient-rated esophageal stricture	s value (relative seriality parameter).	None	1a	AUC, Pearson's χ^2 test	[57]
Grade ≥3 severe late esophagus toxicity	Esophagus EUD, V76.7 Gy to the esophagus	None	1a	None	[58]

was carried out by the same group who had previously developed the model [47], where Pearson's R^2 statistics was used as a measure of overall model performance.

To summarize, out of 7 articles on hypothyroidism, 4 were dedicated to model development (TRIPOD type 1a or 1b), 1 on model development and validation in the same article (TRIPOD type 3) and 2 on model validation only (TRIPOD types 4a & 4b).

Esophagitis

A total of 10 articles (9 models) were reviewed for esophagitis, summarized in Table 6. Details of the articles are provided in Supplementary Table S6.

Quantitative dose–response models of esophagus toxicity are mainly developed for non-small cell lung cancer (NSCLC) patients. Mean dose to esophagus and concurrent chemotherapy were recognized as the most highly reported dosimetric and clinical prognostic factors respectively for esophagitis incidence.

Out of 10 articles, 9 reported on model development only and were assigned TRIPOD type1a or 1b, six of which, all with TRIPOD type 1a, did not provide any performance measures. One type 1a study [57], however, provided discrimination by AUC. Two articles with TRIPOD type 1b provided some performance measures: Huang et al. [52] reported discrimination (AUC) and a calibration plot; Wijsman [50] reported all three measures of model performance using Nagelkerke's R^2, calibration plot coupled with Hosmer-Lemeshow test and AUC. While the model's external validation by Dankers et al. [51] only performed discriminative performance by AUC.

In sum, from 10 total articles, 9 were dedicated to model development only (TRIPOD types 1a & 1b) while only 1 TRIPOD type 4a study was observed.

Discussion

A comprehensive literature review on NTCP models relevant to radiation-induced HNC toxicities was performed. Our review demonstrated scarcity of external validation studies, specifically by independent investigators. The purpose of a prediction model is by definition to provide reliable outcome predictions. To this end, external validation is of high importance, since it is very probable to develop a model which performs well in the training dataset but fails when applied to out-sample data. It is an essential step for generalizability, transportability and causality appraisal of the predicted models in an independent patient cohort.

In this study we utilized the TRIPOD statement as a way to quantitatively analyze the prediction strength of the models. Based on our analysis, 80% of articles are dedicated to model development only, from which only 13% have been externally validated. Two independent external validation studies were observed; one of which validated five models developed for photon radiotherapy on a population of proton therapy patients. The second article was

dedicated to independent external validation of a multivariate model for physician-rated dysphagia using patients from the DAHANCA19 trial as the validation cohort. It should be noted that if the datasets used for external validation are collected by the same researchers who developed the model, this external validation is still not considered as fully independent validation. Moreover, it should be noted that even a single external validation does not label a model as 'validated': the more frequently a model is externally validated and the more diverse are the validation cohort settings (e.g. patients treated with a different technique, at a different hospital or a different time point, etc.) the more assurance can be gained regarding model's generalizability. We would like to mention a type of indirect model validation, i.e., examination of NTCP curves for different clinically-relevant assessments of a specific toxicity. This type of model validation for late dysphagia was studied by Eisbruch et al. [59] and demonstrated that objective assessment of dysphagia, either as Videofluoroscopy Summary Score or Increase in Aspirations both for pharyngeal constrictor mean dose or glottis and supraglottic larynx, led to the same NTCP. The same result was gained for different patient-reported outcomes.

During our analysis, we were also concerned about the usage of performance measures in each article. Based on our analysis, 13% of the articles reported all performance measures (overall model performance, calibration & discrimination). For the sake of transparent reporting, it is highly recommended that authors include performance metrics, specifically discrimination and calibration, in their final report. Transparent reporting would also make the process of independent validation more straightforward; because direct comparison of the discriminative and calibration measures between validation and training cohorts, and finally the external validation of the model, becomes feasible.

The lack of external and independent validation of NTCP models of the head and neck represents also a missed opportunity for the larger radiation oncology community: validated and accurate prediction models could be incorporated as a tool in the clinic and help inform decisions and choices made by the planning team, e.g. on OAR sparing vs PTV coverage. However, several independent external validation studies would be a prerequisite to such developments.

In the context of clinical trials, such NTCP models would also be extremely useful in quality assurance of patient plans. Unacceptable toxicity risk thresholds could be implemented in Radiation Therapy Quality Assurance (RTQA) guidelines to supplement QUANTEC- or consensus-defined dose constraints. The feasibility of such implementation is currently being investigated by our group at the EORTC by validating some of the cited models using an independent patient dataset from the EORTC HNCG-ROG 1219 DAHANCA trial. The final validated models could be selected prospectively as outcome predictors for patients recruited in future EORTC trials and systematically used as a QA tool.

In order to gain assurance on a model's generalizability and clinical usefulness, external validation studies would be highly welcomed by other independent groups as well, to warrant the diversity of patient and treatment-related characteristics and to the benefit of all potential users of NTCP models.

Conflicts of interest

None declared

Acknowledgements

Marjan Sharabiani's work as a Fellow at EORTC Headquarters was supported by a grant from EORTC Cancer Research Fund.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2020.02.013.

References

[1] Brouwer CL, Steenbakkers RJ, Bourhis J, Budach W, Grau C, Grégoire V, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiother Oncol 2015;117:83–90.
[2] Grégoire V, Levendag P, Ang KK, Bernier J, Braaksma M, Budach V, et al. CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC, RTOG consensus guidelines. Radiother Oncol 2003;69:227–36.
[3] Vorwerk H, Hess CF. Guidelines for delineation of lymphatic clinical target volumes for high conformal radiotherapy. Head Neck Region 2011;6.
[4] Langendijk JA, Doornaert P, Rietveld DH, Verdonck-de Leeuw IM, René Leemans C, Slotman BJ. A predictive model for swallowing dysfunction after curative radiotherapy in head and neck cancer. Radiother Oncol 2009;90:189–95.
[5] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38.
[6] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. Eur Urol 2015;67:1142–51.
[7] Brodin NP, Kabarriti R, Garg MK, Guha C, Tomé WA. Systematic review of normal tissue complication models relevant to standard fractionation radiation therapy of the head and neck region published after the QUANTEC reports. Int J Radiat Oncol Biol Phys 2018;100:391–407.
[8] Blanchard P, Wong AJ, Gunn GB, Garden AS, Mohamed AS, et al. Toward a model-based patient selection strategy for proton therapy: External validation of photon-derived normal tissue complication probability models in a head and neck proton therapy cohort. Radiother Oncol 2016;121:381–6.
[9] Hansen CR, Friborg J, Jensen K, Samsøe E, Johnsen L, Zukauskaite R, et al. NTCP model validation method for DAHANCA patient selection of protons versus photons in head and neck cancer radiotherapy. Acta Oncol 2019;58:1410–5.
[10] Chen WC, Lai CH, Lee TF, Hung CH, Liu KC, et al. Scintigraphic assessment of salivary function after intensity-modulated radiotherapy for head and neck cancer: Correlations with parotid dose and quality of life. Oral Oncol 2013;49:42–8.
[11] Lee TF, Chao PJ, Wang HY, Hsu HC, Chang PS, Chen WC. Normal tissue complication probability model parameter estimation for xerostomia in head and neck cancer patients based on scintigraphy and quality of life assessments. BMC Cancer 2012;12:12.
[12] Beetz I, Schilstra C, Burlage FR, Koken PW, Doornaert P, Bijl HP, et al. Development of NTCP models for head and neck cancer patients treated with three-dimensional conformal radiotherapy for xerostomia and sticky saliva: the role of dosimetric and clinical factors. Radiother Oncol 2012;105:86–93.
[13] Beetz I, Schilstra C, Van Luijk P, Christianen ME, Doornaert P, Bijl HP, et al. External validation of three dimensional conformal radiotherapy based NTCP models for patient-rated xerostomia and sticky saliva among patients treated with intensity modulated radiotherapy. Radiother Oncol 2012;105:94–100.
[14] Beetz I, Schilstra C, Van Der Schaaf A, Van Den Heuvel ER, Doornaert P, Van Luijk P, et al. NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: the role of dosimetric and clinical factors. Radiother Oncol 2012;105:101–6.
[15] Lee TF, Chao PJ, Ting HM, Chang L, Huang YJ, Wu JM, et al. Using multivariate regression model with least absolute shrinkage and selection operator (LASSO) to predict the incidence of xerostomia after intensity-modulated radiotherapy for head and neck cancer. PLoS ONE 2014;9.
[16] Lee TF, Liou MH, Huang YJ, Chao PJ, Ting HM, Lee HY, et al. LASSO NTCP predictors for the incidence of xerostomia in patients with head and neck squamous cell carcinoma and nasopharyngeal carcinoma. Sci Rep 2014;4.
[17] Moiseenko V, Wu J, Hovan A, Saleh Z, Apte A, Deasy JO, et al. Treatment planning constraints to avoid xerostomia in head-and-neck radiotherapy: an independent test of QUANTEC criteria using a prospectively collected dataset. Int J Radiat Oncol Biol Phys 2012;82:1108–14.
[18] Houweling AC, Philippens ME, Dijkema T, Roesink JM, Terhaard CH, Schilstra C, et al. A comparison of dose-response models for the parotid gland in a large group of head-and-neck cancer patients. Int J Radiat Oncol Biol Phys 2010;76:1259–65.
[19] Gabryś HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Parotid gland mean dose as a xerostomia predictor in low-dose domains. Acta Oncol 2017;56:1197–203.
[20] Marzi S, Iaccarino G, Pasciuti K, Soriani A, Benassi M, Arcangeli G, et al. Analysis of salivary flow and dose-volume modeling of complication incidence in patients with head-and-neck cancer receiving intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys 2009;73:1252–9.
[21] Miah AB, Gulliford SL, Clark CH, Bhide SA, Zaidi SH, Newbold KL, et al. Dose–response analysis of parotid gland function: What is the best measure of xerostomia? Radiother Oncol 2013;106:341–5.

[22] Roesink M, Moerland MA, Battermann JJ, Hordijk GJ, Terhaard CH. Qantitative dose-volume response analysis of changes in parotid gland function after radiotheraphy in the head-and-neck region. Int J Radiat Oncol Biol Phys 2001;51:938–46.
[23] Dijkema T, Terhaard CHJ, Roesink JM, Braam PM, van Gils CH, Moerland MA, et al. Large cohort dose-volume response analysis of parotid gland function after radiotherapy: intensity-modulated versus conventional radiotherapy. Int J Radiat Oncol Biol Phys 2008;72:1101–9.
[24] Lee TF, Liou MH, Ting HM, Chang L, Lee HY, Wan Leung S, et al. Patient- and therapy-related factors associated with the incidence of xerostomia in nasopharyngeal carcinoma patients receiving parotid-sparing helical tomotherapy. Sci Rep 2015;5.
[25] Mavroidis P, Price A, Fried D, Kostich M, Amdur R, Mendenhall W, et al. Dose–volume toxicity modeling for de-intensified chemo-radiation therapy for HPV-positive oropharynx cancer. Radiother Oncol 2017;124:240–7.
[26] Dijkema T, Raaijmakers CP, Ten Haken RK, Roesink JM, Braam PM, Houweling AC, et al. Parotid gland function after radiotherapy: the combined Michigan and Utrecht experience. Int J Radiat Oncol Biol Phys 2010;78:449–53.
[27] Christianen ME, Schilstra C, Beetz I, Muijs CT, Chouvalova O, Burlage FR, Doornaert P, et al. Predictive modelling for swallowing dysfunction after primary (chemo)radiation: results of a prospective observational study. Radiother Oncol 2012;105:107–14.
[28] Christianen ME, Van Der Schaaf A, Van Der Laan HP, Verdonck-De Leeuw IM, Doornaert P, Chouvalova O, et al. Swallowing sparing intensity modulated radiotherapy (SW-IMRT) in head and neck cancer: clinical validation according to the model-based approach. Radiother Oncol 2016;118:298–303.
[29] Dean J, Wong K, Gay H, Welsh L, Jones A-B, Schick U, et al. Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. Clin Transl Radiat Oncol 2018;8:27–39.
[30] Bhide SA, Gulliford S, Schick U, Miah A, Zaidi S, Newbold K, et al. Dose–response analysis of acute oral mucositis and pharyngeal dysphagia in patients receiving induction chemotherapy followed by concomitant chemo-IMRT for head and neck cancer. Radiother Oncol 2012;103:88–91.
[31] Alterio D, Gerardi MA, Cella L, Spoto R, Zurlo V, Sabbatini A, et al. Strahleninduzierte akute Dysphagie: Prospektive Beobachtungsstudie an 42 Kopf-Hals-Malignompatienten. Strahlenther Onkol 2017;193:971–81.
[32] Tsai CJ, Jackson A, Setton J, Riaz N, McBride S, Leeman J, et al. Modeling dose response for late dysphagia in patients with head and neck cancer in the modern era of definitive chemoradiation. JCO Clin Cancer Informatics 2017;1:1–7.
[33] Wopken K, Bijl HP, Van Der Schaaf A, Van Der Laan HP, Chouvalova O, Steenbakkers RJ, et al. Development of a multivariable normal tissue complication probability (NTCP) model for tube feeding dependence after curative radiotherapy/chemo-radiotherapy in head and neck cancer. Radiother Oncol 2014;113:95–101.
[34] Kanayama N, Kierkels RG, van der Schaaf A, Steenbakkers RJ, Yoshioka Y, Nishiyama K, et al. External validation of a multifactorial normal tissue complication probability model for tube feeding dependence at 6 months after definitive radiotherapy for head and neck cancer. Radiother Oncol 2018;129:403–8.
[35] Wopken K, Bijl HP, van der Schaaf A, Christianen ME, Chouvalova O, Oosting SF, et al. Development and validation of a prediction model for tube feeding dependence after curative (chemo-) radiation in head and neck cancer. PLoS ONE 2014;9:e94879.
[36] Dean JA, Wong KH, Welsh LC, Jones A-B, Schick U, Newbold KL, et al. Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy. Radiother Oncol 2016;120:21–7.
[37] Dean JA, Welsh LC, Wong KH, Aleksic A, Dunne E, Islam MR, et al. Normal Tissue Complication Probability (NTCP) modelling of severe acute mucositis using a novel oral mucosal surface organ at risk. Clin Oncol 2017;29:263–73.
[38] Strigari L, Pedicini P, D'Andrea M, Pinnarò P, Marucci L, Giordano C, et al. A new model for predicting acute mucosal toxicity in head-and-neck cancer patients undergoing radiotherapy with altered schedules. Int J Radiat Oncol Biol Phys 2012;83:e697–702.
[39] Lee T-F, Yeh S-A, Chao P-J, Chang L, Chiu C-L, Ting H-M, et al. Normal tissue complication probability modeling for cochlea constraints to avoid causing tinnitus after head-and-neck intensity-modulated radiation therapy. Radiat Oncol 2015;10.
[40] Cheraghi S, Nikoofar A, Bakhshandeh M, Khoei S, Farahani S, Abdollahi H, et al. Normal tissue complication probability modeling of radiation-induced sensorineural hearing loss after head-and-neck radiation therapy. Int J Radiat Biol 2017;93:1327–33.
[41] De Marzi L, Feuvret L, Boulé T, Habrand J-L, Martin F, Calugaru V, et al. Use of gEUD for predicting ear and pituitary gland damage following proton and photon radiation therapy. Br J Radiol 2015;88. 20140413.
[42] Mosleh-Shirazi MA, Amraee A, Mohaghegh F. Dose-response relationship and normal-tissue complication probability of conductive hearing loss in patients undergoing head-and-neck or cranial radiotherapy: a prospective study including 70 ears. Physica Med 2019;61:64–9.
[43] Bakhshandeh M, Hashemi B, Mahdavi SRM, Nikoofar A, Vasheghani M, Kazemnejad A. Normal tissue complication probability modeling of radiation-induced hypothyroidism after head-and-neck radiation therapy. Int J Radiat Oncol Biol Phys 2013;85:514–21.
[44] Boomsma MJ, Bijl HP, Christianen ME, Beetz I, Chouvalova O, Steenbakkers RJ, et al. A prospective cohort study on radiation-induced hypothyroidism: development of an NTCP model. Int J Radiat Oncol Biol Phys 2012;84:e351–6.
[45] Cella L, Liuzzi R, Conson M, D'Avino V, Salvatore M, Pacelli R. Development of multivariate NTCP models for radiation-induced hypothyroidism: a comparative analysis. Radiat Oncol 2012;7.
[46] Rønjom MF, Brink C, Bentzen SM, Hegedüs L, Overgaard J, Johansen J. Hypothyroidism after primary radiotherapy for head and neck squamous cell carcinoma: normal tissue complication probability modeling with latent time correction. Radiother Oncol 2013;109:317–22.
[47] Rønjom MF, Brink C, Bentzen SM, Hegedüs L, Overgaard J, Petersen JB, et al. External validation of a normal tissue complication probability model for radiation-induced hypothyroidism in an independent cohort. Acta Oncol 2015;54:1301–9.
[48] Luo R, Wu VW, He B, Gao X, Xu Z, Wang D, et al. Development of a normal tissue complication probability (NTCP) model for radiation-induced hypothyroidism in nasopharyngeal carcinoma patients. BMC Cancer 2018;18.
[49] Kwint M, Uyterlinde W, Nijkamp J, Chen C, de Bois J, Sonke J-J, et al. Acute esophagus toxicity in lung cancer patients after intensity modulated radiation therapy and concurrent chemotherapy. Int J Radiat Oncol Biol Phys 2012;84:e223–8.
[50] Wijsman R, Dankers F, Troost EG, Hoffmann AL, van der Heijden EH, de Geus-Oei L-F, et al. Multivariable normal-tissue complication modeling of acute esophageal toxicity in advanced stage non-small cell lung cancer patients treated with intensity-modulated (chemo-)radiotherapy. Radiother Oncol 2015;117:49–54.
[51] Dankers FJ, Wijsman R, Troost EG, Tissing-Tan CJ, Kwint MH, Belderbos J, et al. External validation of an NTCP model for acute esophageal toxicity in locally advanced NSCLC patients treated with intensity-modulated (chemo-) radiotherapy. Radiother Oncol 2018;129:249–56.
[52] Huang EX, Bradley JD, El Naqa I, Hope AJ, Lindsay PE, Bosch WR, et al. Modeling the risk of radiation-induced acute esophagitis for combined Washington University and RTOG trial 93–11 lung cancer patients. Int J Radiat Oncol Biol Phys 2012;82:1674–9.
[53] Zhu J, Zhang ZC, Li BS, Liu M, Yin Y, Yu JM, et al. Analysis of acute radiation-induced esophagitis in non-small-cell lung cancer patients using the Lyman NTCP model. Radiother Oncol 2010;97:449–54.
[54] Chapet O, Kong FM, Lee JS, Hayman JA, Ten Haken RK. Normal tissue complication probability modeling for acute esophagitis in patients treated with conformal radiation therapy for non-small cell lung cancer. Radiother Oncol 2005;77:176–81.
[55] Belderbos J, Heemsbergen W, Hoogeman M, Pengel K, Rossi M, Lebesque J. Acute esophageal toxicity in non-small cell lung cancer patients after high dose conformal radiotherapy. Radiother Oncol 2005;75:157–64.
[56] Zehentmayr F, Söhn M, Exeli AK, Wurstbauer K, Tröller A, Deutschmann H, et al. Normal tissue complication models for clinically relevant acute esophagitis (≥ grade 2) in patients treated with dose differentiated accelerated radiotherapy (DART-bid). Radiat Oncol 2015;10.
[57] Mavroidis P, Laurell G, Kraepelien T, Fernberg JO, Lind BK, Brahme A. Determination and clinical verification of dose-response parameters for esophageal stricture from head and neck radiotherapy. Acta Oncol 2003;42:865–81.
[58] Chen C, Uyterlinde W, Sonke JJ, de Bois J, van den Heuvel M, Belderbos J. Severe late esophagus toxicity in NSCLC patients treated with IMRT and concurrent chemotherapy. Radiother Oncol 2013;108:337–41.
[59] Eisbruch A, Kim HM, Feng FY, Lyden TH, Haxer MJ, Feng M, et al. Chemo-IMRT of oropharyngeal cancer aiming to reduce dysphagia: swallowing organs late complication probabilities and dosimetric correlates. Int J Radiat Oncol Biol Phys 2011;81:e93–9.